# DRAGON SYSTEMS' 1998 BROADCAST NEWS TRANSCRIPTION SYSTEM

*Steven Wegmann, Puming Zhan, Ira Carp, Michael Newman, Jon Yamron, and Larry Gillick*

Dragon Systems, Inc.
320 Nevada Street, Newton, MA  02460

## ABSTRACT

In this paper we shall describe key improvements to Dragon's Broadcast News Transcription System, which include: the addition of a speaker-change detection algorithm to our preprocessing subsystem, a new diagonalizing transformation trained using semi-tied covariances, and the addition of probabilities on pronunciations. This new transcription system yields a word error rate of 15.2% on the 1997 evaluation test data, and 14.5% on the 1998 evaluation test data.

## 1. INTRODUCTION

We have made substantial progress on the Broadcast News Transcription task since the 1997 Broadcast News evaluation ([1]), but we have done so without major architectural changes to our system. Instead, we attended to the details of implementation, which paid off handsomely.

In the postmortem of last year's evaluation we identified our very primitive silence-based preprocessing system as a fertile ground for potential improvements, and indeed we made substantial improvements by including a speaker detection algorithm ([3]), and cleaning up the basic system. In the acoustic modelling arena, we replaced our standard IMELDA ([2]) transformation with a generalization, which we are calling a diagonalizing transformation. Nearly 5 absolute percentage points of our overall improvement comes from these two changes.

Since the general structure of our new transcription system is almost identical to last year's system, which is described in [1], in this paper we shall focus on the improvements that we have made.

## 2. EXPERIMENTS & ANALYSIS

### 2.1 Preliminaries

In the following sections we shall present a series of experiments designed to uncover how much improvement we get from the changes that we made to this year's system. We shall start from our complete 1998 evaluation system and gradually remove our improvements. As a consequence, in all of the experiments, unless otherwise noted, the language model (LM) is the 1998 evaluation trigram LM (which is described in section 2.4). All results reported are from the second, adapted, recognition pass (we are using unsupervised rapid adaptation with one transformation [4], [5]). All of the acoustic models were trained from warped data and all of the test data are warped (as in [1]). Finally, all experimental results report word error rate (WER) as measured on the 1997 evaluation test set, broken out using the standard focus conditions.

Table 1 compares our official 1997 submission (Old) with our new 1998 evaluation system. Recall that subsequent to the 1997 evaluation, we discovered that our recognizer settings were too tight, and after retuning (on a different test set), our system error rate went down to 21.4 ([1]).

|  | Old | New |
|---|---|---|
| F0 | 13.9 | 9.5 |
| F1 | 23.4 | 15.0 |
| F2 | 31.1 | 20.4 |
| F3 | 34.9 | 21.5 |
| F4 | 26.5 | 20.2 |
| F5 | 19.0 | 18.8 |
| FX | 43.9 | 32.6 |
| Total | 23.1 | 15.2 |

**Table 1:** Comparison of 1997 and 1998 systems.

### 2.2 ROVER

The preprocessing system's job is to chop the input broadcast stream into reasonably sized homogeneous segments which are clustered into speaker-like groups for the purposes of warping and adaptation (see [1] for a description of our preprocessing system).

|  | Sys1 | Sys2 | ROVER |
|---|---|---|---|
| F0 | 9.9 | 10.2 | 9.5 |
| F1 | 15.5 | 15.7 | 15.0 |
| F2 | 20.9 | 21.4 | 20.4 |
| F3 | 21.6 | 23.5 | 21.5 |
| F4 | 21.3 | 20.6 | 20.2 |
| F5 | 20.1 | 19.2 | 18.8 |
| FX | 33.2 | 33.5 | 32.6 |
| Total | 15.7 | 16.0 | 15.2 |

**Table 2:** Improvement due to ROVER.

Towards the end of our development effort, we noticed that our preprocessing system was still responsible for about 1 percentage point of errors, instead of the expected 0.5 point. In the course of exploring various causes for this, we discovered that we could cut our error rate by about 0.5 percentage points by using ROVER ([6]) to combine the output of two complete systems. These systems differed only in how the initial segments were created, but these slightly different initial segments led to slightly different clusters, which in turn led to slightly different warps for the clusters, etc.

In Table 2 we compare the performance of the combined system with the two input systems, labeled ROVER, Sys1, and Sys2 respectively. Sys1 was based on segments determined by a coarse recognition pass using left diphone models without crossword co-articulation, while Sys2 used segments determined by a coarse recognition pass using standard triphone models with crossword co-articulation. By using ROVER we were able to recover the errant 0.5 point. In the following sections, we will be using the automatic segments produced by Sys1.

## 2.3 Probabilities on Pronunciations

Following in the steps of Dragon's SWITCHBOARD (SWB) effort ([7]), we used bigram probabilities for alternate pronunciations in our evaluation system. These probabilities were trained from forced alignments of the acoustic training transcripts. Our evaluation lexicon has 57K words with 62K pronunciations. In Table 3, we compare performance of systems with bigram, unigram, and no pronunciation probabilitie (i.e. all alternate pronuciations are equiprobable for a given word), labeled Bigram, Unigram, and None respectively.

| | None | Unigram | Bigram |
|---|---|---|---|
| F0 | 10.3 | 10.2 | 10.0 |
| F1 | 16.5 | 16.4 | 15.9 |
| F2 | 21.7 | 21.4 | 21.1 |
| F3 | 24.0 | 22.0 | 22.1 |
| F4 | 21.2 | 21.8 | 21.1 |
| F5 | 21.7 | 20.4 | 20.2 |
| FX | 32.3 | 32.6 | 32.2 |
| Total | 16.3 | 16.1 | 15.8 |

**Table 3:** Effect of adding unigram and bigram probabilities on alternate pronunciations.

We saw an improvement of about a percentage point when we tried using unigram probabilities in early development work on this corpus, which agreed with experiments on the SWB corpus ([7]). Now we are getting very little from the unigram probabilities, 0.2 percentage points, and only 0.5 percentage when we use bigram probabilities. We shall be investigating this more carefully in the future.

## 2.4 Language Modelling

Like last year, we used a three-way interpolated language model. The three components were backoff trigram language models were trained from about 770 million words of text. The first component was trained from the Broadcast News acoustic training transcripts plus the 1995 Marketplace development transcripts (1.6 million words, about double what was available last year). The second component was the same as last year, that is, it was trained from the Broadcast News language model training corpus (130 million words). The third component was trained from the 1995 Hub4 and Hub3 newswire texts plus all of the allowable texts from the LDC's North American News Text Corpus supplement (640 million words). Last year we supplemented this component with commercially available newspaper texts, because the LDC's newswire supplement was not available.

These language models share a 57K vocabulary constructed from the combined training sources. This 57K set resulted in an OOV rate of 0.6% on the 1998 evaluation test.

| | Old | New |
|---|---|---|
| F0 | 10.2 | 10.3 |
| F1 | 16.5 | 16.5 |
| F2 | 22.2 | 21.7 |
| F3 | 24.3 | 24.0 |
| F4 | 21.3 | 21.2 |
| F5 | 20.3 | 21.7 |
| FX | 33.9 | 32.3 |
| Total | 16.4 | 16.3 |

**Table 4:** Comparison of 1997 and 1998 language models.

In Table 4 we compare the performance of last year's LM with this year's LM on the 1997 evaluation test. We are not using probabilities on pronunciations with either of these experiments, so the New column corresponds to the None column in Table 3. There is no significant difference between the performance of these two language models on the 1997 evaluation data. So adding the new acoustic training texts and the supplementary newswire data to our language model training has balanced the loss of the newspaper texts.

## 2.5 Acoustic Training

This year there were 140 hours of acoustic training data compared to the 70 hours available last year. We get nearly a percentage point improvement from adding the new data, as Table 5 shows (both these results use the old, 1997 evaluation LM, so the column labeled "140 hrs" corresponds to the "Old" column in Table 4).

Table 6 breaks this improvement out by gender. This result is somewhat surprising since, given that the females make up about a third of the data in the training corpus we might expect that the 70-hour models are starved for female data. If that were the case, then we would expect that the female test speakers would have improved more than the males. The fact that these models are warped and GD may explain why the 70 hour models were not starved for female data. Both techniques have worked together to reduce the variability between males and females.

| | 70 hrs | 140 hrs |
|---|---|---|
| F0 | 10.7 | 10.2 |
| F1 | 17.1 | 16.5 |
| F2 | 22.7 | 22.2 |
| F3 | 27.2 | 24.3 |
| F4 | 22.6 | 21.3 |
| F5 | 26.3 | 20.3 |
| FX | 36.2 | 33.9 |
| Total | 17.3 | 16.4 |

**Table 5:** Effect of training size.

| | Male | Female |
|---|---|---|
| 140 hrs | 17.1 | 15.3 |
| 70 hrs | 18.2 | 15.9 |

**Table 6**: Effect of training size broken by gender.

## 2.6 Gender Dependent Modelling

We used gender dependent (GD) acoustic models for first time in conjunction with warping ([8]). The gender dependent models were trained by adapting gender independent (GI) models to the gender specific data ([9]). Table 7 presents first pass and second pass, adapted recognition results using GD and GI models. Before adaptation it is an overall win to use GD models, but mainly for the females, while after adaptation it is only a win for the females. (The GD row in Table 7 corresponds to "None" column in Table 3.)

| | Total | | Male | | Female | |
|---|---|---|---|---|---|---|
| | | Adapt | | Adapt | | Adapt |
| GD | 18.7 | 16.3 | 20.0 | 17.0 | 16.7 | 15.1 |
| GI | 19.3 | 16.4 | 20.2 | 16.9 | 17.9 | 15.5 |

**Table 7:** Effect of GD modelling.

## 2.7 Preprocessing

We have made substantial improvements to our preprocessing system, which have in turn led to significantly fewer recognition errors. We concentrated our efforts on improving the quality of the segments that we were producing. Like last year, we produce our segments by looking for sufficiently long silence regions in the output of a coarse recognition pass.

The biggest change that we made to our segment generation system was the addition of a speaker change detection algorithm, which is described in [3] and [8]. We did this to increase the purity of the segments that we are producing, which should in turn lead to more homogeneous clusters, which should lead to better warp selection and adaptation, since they both take place on a per cluster basis. This year we are refining a segment if a speaker change is hypothesized within the segment. In Table 8 we compare Sys1, and uses the speaker change detection algorithm, with "No SCD" which is Sys1 minus the speaker change detection algorithm. Recall, from section 2.2, that Sys1's segments were based on the output of a coarse recognition pass using left diphone models without crossword co-articulation.

| | No SCD | Sys1 |
|---|---|---|
| F0 | 10.5 | 10.3 |
| F1 | 17.2 | 16.5 |
| F2 | 23.3 | 21.7 |
| F3 | 24.0 | 24.0 |
| F4 | 21.5 | 21.2 |
| F5 | 20.7 | 21.7 |
| FX | 33.3 | 32.3 |
| Total | 16.8 | 16.3 |

**Table 8:** Effect of speaker change detection.

We also made several seemingly minor changes, which led to big overall improvements when put together. We base our segments on the output of a word recognizer with more carefully tuned settings than last year's phoneme recognizer. These two changes were an attempt to prevent segment boundaries from being placed in the middle of words. We also assign a gender label to each segment and then cluster separately for each gender.

| | Old | New |
|---|---|---|
| F0 | 10.8 | 10.4 |
| F1 | 19.3 | 16.6 |
| F2 | 25.5 | 22.3 |
| F3 | 28.3 | 23.9 |
| F4 | 23.4 | 21.3 |
| F5 | 22.7 | 19.7 |
| FX | 33.9 | 32.5 |
| Total | 18.2 | 16.4 |

**Table 9:** Improvement due to preprocessing improvements.

How much better is our current preprocessing system than last year's? Our old system did not include gender detection, so we make this comparison with gender independent acoustic models (we are using this year's LM and the "New" system includes speaker change detection). Table 9 shows that we have made a 1.8 percentage point improvement just from better preprocessing.

## 2.8 Diagonalizing Transformations

In the past we have done our acoustic modelling in a 24 dimensional space obtained by first applying an IMELDA transformation to our 36 dimensional feature space (12 PLP-based cepstra, along with their first and second differences), and then projecting down to a 24 dimensional subspace. A large share of the improvement that we made this year comes from replacing the IMELDA transformation with a generalization, which we call a diagonalizing transformation, and using the resulting 36 dimensional feature space, i.e. without further projection.

The motivating idea behind this new transformation is simple: since we assume a diagonal covariance in the multivariate gaussians used in our acoustic models, we should seek a representation of acoustic space that most closely agrees with this assumption. This technique is due to Gales [10] and Kumar [11], and was first applied to the Broadcast News corpus by Gopinath [12].

How much do we get from this new technique? In early development experiments, where we used a small bigram language model, we have seen about a 2.5 percentage point improvement after adaptation (see [8]). We can get a more accurate reading of how much better these models are by comparing the performance of last year's acoustic and language models to GI versions of this year's acoustic and language models (without probabilities on pronunciations) on the 1997 evaluation test using last year's preprocessing system. Table 10 displays the results, which require some interpretation. As we saw in section 2.4 the difference in language models needn't concern us, but section 2.5 showed that the extra 70 hours of training data that the new models used gave a 0.9 point improvement. So if we forget that last year's acoustic models were also trained using SAT with WSJ and WSJCAM0 data, then we get a lower bound of about 2.3 percentage points from this new transformation.

| | Old | New |
|---|---|---|
| F0 | 12.9 | 10.8 |
| F1 | 22.2 | 19.3 |
| F2 | 30.2 | 25.5 |
| F3 | 33.4 | 28.3 |
| F4 | 27.7 | 23.4 |
| F5 | 18.4 | 22.7 |
| FX | 43.0 | 33.9 |
| Total | 21.4 | 18.2 |

**Table 10:** Comparison of 1997 and 1998 models.

It is also interesting to note how large the improvement in the degraded conditions, namely, the low bandwidth (F2), music (F3), noise (F4), and the catchall (FX) categories. In fact, we were planning to use separate, low bandwidth acoustic models to decode the low bandwidth data, but we could not improve on the performance shown in Table 10.

## 3. FUTURE WORK

Thus far we have concentrated on techniques that improve the error rate in all conditions. We shall begin to explore techniques that work particularly well in degraded acoustical conditions. For example, our preprocessing system tends to fall apart in degraded conditions. We also need to work on our language model. For example, we are currently exploring replacing our trigram language models with four grams.

## Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency. The views and findings contained in this material are those of the authors and do not necessarily reflect the position or policy of the U.S. Government and no official endorsement should be inferred.

## REFERENCES

[1] Wegmann, S., et al. "Dragon Systems' 1997 Broadcast News Transcription System". *Proc. Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 60-65. Feb., 1998.

[2] Hunt, M. et al., "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," *Proc. ICASSP-91*, Toronto, May 1991.

[3] Zhan, P., Wegmann, S., and Gillick, L., "Improvements to Dragon Systems' 1998 Mandarin Transcription System*", these Proceedings*.

[4] Leggetter, C. and Woodland, P., "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression," *Proc. ICSLP'94*, Yokohama, September 1994.

[5] Nagesha, V. and Gillick, L., "Studies in Transformation Based Adaptation," *Proc. ICASSP-97*, Munich, April 1997.

[6] Fiscus, J. "A Post-Processing System to Yield Reduced Word Error Rates". *Proc. IEEE ASRU Workshop*, Santa Barbara, pp. 347-352, Dec. 1997.

[7] Peskin, B. et al., "Improvements in Recognition of Conversational Telephone Speech", *Proc. ICASSP-99*, Phoenix, March 1999.

[8] Wegmann, S., Zhan, P., and Gillick, L., "Progress in Broadcast News Transcription at Dragon Systems", *Proc. ICASSP-99*, Phoenix, March 1999.

[9] Woodland, P., et al. "The 1997 HTK Broadcast News Transcription System". *Proc. Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, pp. 41-48. Feb., 1998.

[10] Gales, M. "Semi-tied Covariance Matrices*", Proc. ICASSP-98*, Seattle, 1998.

[11] Kumar, N. "Investigation of Silicon-Auditory Models and Generalizations of LDA for Improved Speech Recognition". *PhD Thesis*, Johns Hopkins University, 1997.

12] Gopinath, R. "Constrained Maximum Likelihood Modeling with Gaussian Distributions", *Proc. ICASSP-98*, Seattle, 1998.